Abstract

Questions are used in different applications such as Question Answering (QA), Dialog System (DS), and Information Retrieval (IR). Unlike some questions, most inquiries can be too complex to be analysed and processed by machines. As a result, systems are expected to have a good feature extraction and analysis mechanism to linguistically understand these questions.

The retrieval of wrong answers, inaccuracy of IR, and crowding the search space with irrelevant candidate answers are some of the challenges that are caused due to the inability to appropriately process and analyse questions.

Question Classification (QC) aims to solve this issue by extracting the relevant features from the questions and by assigning them to the correct class category. Even though QC has been studied for various languages, it was hardly studied for the Amharic language. This research studies Amharic QC focusing on designing hierarchical question taxonomy, preparing Amharic question dataset by labelling (fine-tuning) the sample questions into their respective classes, and implementing AQC model using Convolutional Neural Network (CNN) which is part of the DL approach.

The AQC uses a multilabel question taxonomy that integrates coarse and fine grain categories. These multilabel classes help us to be more accurate in retrieving answers compared to the flat taxonomy. We constructed the taxonomy by analysing our AQ dataset and also adopting the standard taxonomies that were previously studied. We have prepared the AQs in three forms: Surface, Stemmed, and Lemmatised forms. We train and test these datasets using a word vectorizer trained on surface words after noticing that most interrogative words appear to have a similar form even when they are stemmed and lemmatized.

As a result, we have achieved 97% and 90% training and validation accuracy for Surface AQs. Scoring 15% for the stemmed AQs. Nevertheless, the word2vec model could not represent the lemmatized AQs appropriately.

Keywords: - Amharic Question Classification, Deep Learning, CNN, Fine grain, Coarse grain Hierarchical Taxonomies, Word2vec.